

# Efficient and Expressively Complete XML Query Languages

Pablo Barceló   Leonid Libkin  
University of Toronto

# Motivation

---

- Recent interest on **unranked** trees, mostly due to XML applications.
- Logics tend to talk about paths + navigational properties.
- Clear similarities between:
  - **XML** query languages (XPath, XQuery, ...).
  - **Temporal** logics (LTL, CTL\*, ...)
- We want to investigate these connections.

# Motivation

---

- Logics commonly used in XML applications: fragments of FO and MSO.
- **Problem:** Over strings, **model checking** of FO and MSO is non-elementary [Frick & Grohe, 02].
- **Goal:** Find logic  $\mathcal{L}$  such that over trees:
  - $\mathcal{L} = \text{FO}$  or  $\mathcal{L} = \text{MSO}$ .
  - $\mathcal{L}$  has “good” model checking properties.

# Temporal Logics are Good Candidates

---

- Why?
  - Over **full binary** trees,  $\text{FO} = \text{CTL}^*$  [Hafer & Thomas, '87].
  - Over **infinite binary** trees,  $\text{MSO} = L_\mu$  [Niwinski, '88].
  - Over trees (more generally, DAGs), **model checking** of  $L_\mu$  is  $\mathcal{O}(\|\mathcal{A}\| \cdot \|\phi\|^2)$  [Mateescu, 02]. Furthermore,  $\text{CTL}^*$  is translatable into  $L_\mu$ .
- **Idea:** Extend characterizations to unranked trees.
  - We want to characterize whole FO and MSO, not only its **bisimulation-invariant** fragment (XML applications).



# Logical Definability

---

MSO =  $\mathcal{L}$  for **Boolean queries** if:

- $\forall$  sentence  $\phi \in \text{MSO} \exists \phi' \in \mathcal{L}$  such that

$$T \models \phi \iff (T, \epsilon) \models \phi' .$$

- $\forall \psi \in \mathcal{L} \exists$  sentence  $\psi' \in \text{MSO}$  such that

$$(T, \epsilon) \models \psi \iff T \models \psi' .$$

- FO =  $\mathcal{L}$  defined in the same way.

# Logical Definability

---

MSO =  $\mathcal{L}$  for **unary queries** if:

- $\forall \phi(x) \in \text{MSO} \exists \phi' \in \mathcal{L}$  such that

$$T \models \phi(s) \iff (T, s) \models \phi'.$$

- $\forall \psi \in \mathcal{L} \exists \psi'(x) \in \text{MSO}$  such that

$$(T, s) \models \psi \iff T \models \psi'(s).$$

- FO =  $\mathcal{L}$  defined in the same way.

# $\mu$ -Calculus

---

- Logic  $L_\mu[\prec_{\text{ch}}, \prec_{\text{ns}}]$ :

|           |  |   |
|-----------|--|---|
| $\phi :=$ | $a, a \in \Sigma$                        | label                                   |
|           | $X, X \in Var$                           | variables                               |
|           | $\neg\phi$                               | negation                                |
|           | $\phi \vee \phi$                         | disjunction                             |
|           | $\langle \prec_{\text{ch}} \rangle \phi$ | $\phi$ true in a child                  |
|           | $\langle \prec_{\text{ns}} \rangle \phi$ | $\phi$ true in next sibling             |
|           | $\mu X. \phi(X)$                         | least fixed-point of operator $\phi(X)$ |

- In  $L_\mu[\prec_{\text{ch}}]$  there is no  $\langle \prec_{\text{ns}} \rangle$  constructor.

## MSO vs $L_\mu$ : Boolean Queries

---

- $L_\mu^{\text{full}}$  is extension of  $L_\mu$  with past operators  $\langle \prec_{\text{ch}} \rangle^{-1} \phi$  and  $\langle \prec_{\text{ns}} \rangle^{-1} \phi$ .
- **Theorem:** For Boolean queries,

$$\text{MSO}[\prec_{\text{ch}}, \prec_{\text{ns}}] = L_\mu^{\text{full}}[\prec_{\text{ch}}, \prec_{\text{ns}}].$$

- Very little **past** required. In fact,

$$\text{MSO}[\prec_{\text{ch}}, \prec_{\text{ns}}] = L_\mu[\prec_{\text{fch}}, \prec_{\text{ns}}],$$

where  $s \prec_{\text{fch}} s'$  if  $s'$  is the first child of  $s$  wrt  $\prec_{\text{ns}}$ .

## MSO vs $L_\mu$ : Boolean Queries

---

- **Very good:** A language as expressive as MSO, but complexity **linear** on both size of  $\phi$  and  $\mathcal{A}$ .
- It is to be studied if  $L_\mu$  expresses natural XML properties in a simple way.

# MSO vs $L_\mu$ : Boolean Queries

---

- What about  $\text{MSO}[\prec_{\text{ch}}]$ ?
- $C_\mu$  is  $L_\mu$  extended with  $\langle \prec_{\text{ch}} \rangle^k \phi$ , meaning there are at least  $k$  children where  $\phi$  holds.
- **Theorem:** [Walukiewicz, '02] For Boolean queries,

$$\text{MSO}[\prec_{\text{ch}}] = C_\mu[\prec_{\text{ch}}].$$

## MSO vs $L_\mu$ : Unary Queries

---

- Unary queries in MSO?
- For  $\text{MSO}[\prec_{\text{ch}}, \prec_{\text{ns}}]$  nothing to add:

**Theorem:** For unary queries,  $\text{MSO}[\prec_{\text{ch}}, \prec_{\text{ns}}] = L_\mu^{\text{full}}[\prec_{\text{ch}}, \prec_{\text{ns}}]$ .

- For  $\text{MSO}[\prec_{\text{ch}}]$  we need something new:

**Theorem:** For unary queries,  $\text{MSO}[\prec_{\text{ch}}] = C_\mu[\prec_{\text{ch}}] + \langle \prec_{\text{ch}} \rangle^{-1} \phi$ .

# CTL<sup>\*</sup>

---

- CTL<sup>\*</sup> consists of:

**State formulas**  $:= a, a \in \Sigma \mid \neg\phi \mid \phi \vee \phi \mid \mathbf{E}\psi$

**Path formulas**  $:= \phi \mid \neg\psi \mid \psi \vee \psi \mid \mathbf{X}_{\otimes}\psi \mid \psi\mathbf{U}_{\otimes}\psi'$

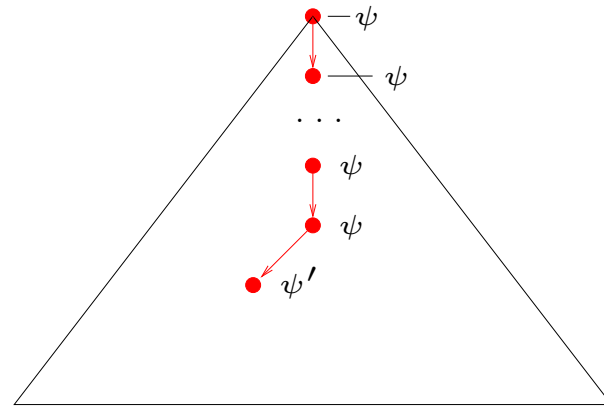
where  $\otimes \in \{\prec_{\text{ch}}, \prec_{\text{ns}}\}$ .

- State formulas are evaluated on elements, and path formulas are evaluated on  $\prec_{\text{ch}}$ -paths or  $\prec_{\text{ns}}$ -paths.

# CTL\*

---

- **Semantics:** given a  $\prec_{\text{ch}}$ - or  $\prec_{\text{ns}}$ -path  $\pi = (s_0, s_1, s_2, \dots)$ :
  - $(T, \pi) \models \mathbf{X}_{\prec_{\text{ch}}} \psi$ , if  $\pi$  is  $\prec_{\text{ch}}$ -path and  $(T, (s_1, s_2, \dots)) \models \psi$  (equivalent for  $\mathbf{X}_{\prec_{\text{ns}}}$ ).
  - $(T, \pi) \models \psi \mathbf{U}_{\prec_{\text{ch}}} \psi'$ ,



- $(T, s) \models \mathbf{E} \psi$ , if for some  $\pi$  starting in  $s$ ,  $(T, \pi) \models \psi$ .

# FO vs CTL<sup>\*</sup>: Boolean Queries

---

- CTL<sub>past</sub><sup>\*</sup> is the extension of CTL<sup>\*</sup> extended with  $\mathbf{X}^{-1}\psi$  (**previous**) and  $\psi \mathbf{S} \psi'$  (**since**).
- A version of CTL<sub>past</sub><sup>\*</sup>: CTL<sub>past</sub><sup>\*</sup>[ $\prec_{\text{ch}} \cup \prec_{\text{ns}}$ ] in which paths refer to the union of relations  $\prec_{\text{ch}}$  and  $\prec_{\text{ns}}$ . Hence, we have **unique** Until and Since operators.
- **Theorem:** For Boolean queries,

$$\text{FO}[\prec_{\text{ch}}^*, \prec_{\text{ns}}^*] = \text{CTL}_{\text{past}}^*[\prec_{\text{ch}}, \prec_{\text{ns}}] = \text{CTL}_{\text{past}}^*[\prec_{\text{ch}} \cup \prec_{\text{ns}}].$$

This is closely related to [Marx, '04].

## FO vs CTL<sup>\*</sup>: Boolean Queries

---

- **Very good (again):** A language as expressive as FO, but complexity **polynomial** on both size of  $\phi$  and  $\mathcal{A}$ .
- **In this case:** CTL<sup>\*</sup> expresses in a very natural way **navigational** properties of XML documents.

**Example:** Formula

$$\mathbf{E} \left( a \mathbf{U}_{\prec_{\text{ch}}} \left( b \wedge \mathbf{F}_{\prec_{\text{ns}}} c \right) \right),$$

expresses that there exists a **descendant** path on which  $a$  holds until ( $b$  holds, and  $c$  holds in a younger sibling).

# FO vs CTL<sup>\*</sup>: Boolean Queries

---

- And FO[ $\prec_{\text{ch}}^*$ ]?
- CTL<sub>count</sub><sup>\*</sup> is CTL<sup>\*</sup>[ $\prec_{\text{ch}}$ ] extended with  $\mathbf{EX}^k\psi$ , meaning there are  $k$  children where  $\psi$  is true.
- **Theorem:** [Moller & Rabinovich, '99] Over Boolean queries,

$$\text{FO}[\prec_{\text{ch}}^*] = \text{CTL}_{\text{count}}^*[\prec_{\text{ch}}].$$

# FO vs CTL<sup>\*</sup>: Unary Queries

---

- Unary queries in FO?

- For  $\text{FO}[\prec_{\text{ch}}^*, \prec_{\text{ns}}^*]$  nothing to add:

**Theorem:** For unary queries,  $\text{FO}[\prec_{\text{ch}}^*, \prec_{\text{ns}}^*] = \text{CTL}_{\text{past}}^*[\prec_{\text{ch}}, \prec_{\text{ns}}]$ .

- For  $\text{FO}[\prec_{\text{ch}}^*]$  we need something new:

**Theorem:** For unary queries,  $\text{FO}[\prec_{\text{ch}}^*] = \text{CTL}_{\text{count}}^*[\prec_{\text{ch}}] + \mathbf{X}^{-1} + \mathbf{S}$ .

## Extensions: Order-invariant MSO

---

- $(\text{MSO}[\prec_{\text{ch}}] + \prec)_{\text{inv}}$ : sentences  $\phi \in \text{MSO}[\prec_{\text{ch}}, \prec]$  such that for any  $\prec_1, \prec_2$ ,

$$(T, \prec_1) \models \phi \iff (T, \prec_2) \models \phi,$$

where  $\prec_1$  and  $\prec_2$  are linear orders.

- **Fact:**  $(\text{MSO}[\prec_{\text{ch}}] + \prec)_{\text{inv}}$  is more expressive than  $\text{MSO}[\prec_{\text{ch}}]$ .
- We could also define  $(\text{MSO}[\prec_{\text{ch}}] + \prec_{\text{ns}})_{\text{inv}}$ , but it has the same expressive power than  $(\text{MSO}[\prec_{\text{ch}}] + \prec)_{\text{inv}}$ .

# Extensions: Order-invariant MSO

---

- $C_{\mu}^{\text{mod}}[\prec_{\text{ch}}]$  is  $C_{\mu}[\prec_{\text{ch}}]$  extended with modulo counting.
- Trivial approach does not work: positive formulas not necessarily **monotone**.

This can be fixed by using techniques in [Muscholl, Schwentick, Seidl & Habermehl, '04].

- **Theorem:**  $(\text{MSO}[\prec_{\text{ch}}] + <)_{\text{inv}} = C_{\mu}^{\text{mod}}[\prec_{\text{ch}}]$ .

Note on proof: By elementary characterization of tree automata for  $(\text{MSO}[\prec_{\text{ch}}] + <)_{\text{inv}}$ . This gives us an elementary proof of:

**Corollary:** [Courcelle, '91]  $(\text{MSO}[\prec_{\text{ch}}] + <)_{\text{inv}} = \text{CMSO}[\prec_{\text{ch}}]$   
(=  $\text{MSO}[\prec_{\text{ch}}] + \text{modulo quantifiers}$ ).

# Extensions: Conjunctive Queries

---

- With temporal logics we can also characterize useful fragments of logics.
- Consider the **fragment** of  $\text{CTL}_{\text{past}}^*[\prec_{\text{ch}}, \prec_{\text{ns}}]$  without negation, and where we only allow:
  - $\text{true } \mathbf{U} \phi$  (**at some point in the future**), and
  - $\text{true } \mathbf{S} \phi$  (**at some point in the past**).

**Theorem:** Over unary queries, this fragment is equivalent to

$$\bigcup \text{CQ}(\prec_{\text{ch}}, \prec_{\text{ch}}^*, \prec_{\text{ns}}, \prec_{\text{ns}}^*).$$

## Further Work

---

- Expressive power of  $(\text{FO}[\prec_{\text{ch}}^*] + \prec_{\text{ns}}^*)_{\text{inv}}$  ?
  - **Conjecture:**  $(\text{FO}[\prec_{\text{ch}}^*] + \prec_{\text{ns}}^*)_{\text{inv}} = \text{FO}[\prec_{\text{ch}}^*]$ .
- Extension of results to  $n$ -ary queries.
  - Our proof methods based on composition are well-suited to deal with them.